Robust Multimodal Person Identification With Limited Training Data

McLaughlin, N., Ji, M., & Crookes, D. (2013). Robust Multimodal Person Identification With Limited Training Data. *IEEE Transactions on Human Machine Systems*, *43*(2), 214 - 224. [6461532]. https://doi.org/10.1109/TSMCC.2012.2227959

Published in:
IEEE Transactions on Human Machine Systems

Document Version:
Peer reviewed version

Robust Multimodal Person Identification with Limited Training Data

Niall McLaughlin, Ji Ming, Danny Crookes

Abstract

This paper presents a novel method of audio-visual feature-level fusion for person identification where both the speech and facial modalities may be corrupted, and there is a lack of prior knowledge about the corruption. Furthermore, we assume there is a limited amount of training data for each modality (e.g., a short training speech segment and a single training facial image for each person). A new multimodal feature representation and a modified cosine similarity are introduced for combining and comparing bimodal features with limited training data as well as vastly differing data rates and feature sizes. Optimal feature selection and multicondition training are used to reduce the mismatch between training and testing, thereby making the system robust to unknown bimodal corruption. Experiments have been carried out on a bimodal data set created from the SPIDRE speaker recognition database and AR face recognition database with variable noise corruption of speech and occlusion in the face images. The system's speaker identification performance on the SPIDRE database, and facial identification performance on the AR database, is comparable with the literature. Combining both modalities using the new method of multimodal fusion leads to significantly improved accuracy over the unimodal systems, even when both modalities have been corrupted. The new method also shows improved identification accuracy compared to the bimodal systems based on multicondition model training or missing-feature decoding alone.

Index Terms

Multimodal fusion, noisy speech, occluded face, robustness, person identification, limited training data

I. Introduction

Biometric person identification becomes a challenging problem when the data used for identification is corrupted. The problem may be further compounded by a shortage of training data in both modalities. By

This work was supported by Intel.

The authors are with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K. (e-mail: nmclaughlin02@qub.ac.uk, j.ming@qub.ac.uk, d.crookes@qub.ac.uk).

combining information from multiple modalities it is possible for a multi-modal biometric system to both increase identification accuracy [1] and overcome the limitations of corruption in a single modality [2], [3], [4], [5], [6]. We consider the fusion of speech and facial data for person identification, assuming corruption of both modalities, i.e., nonstationary full-band corruption of speech and partial occlusion of the face. In addition we assume there is only a limited amount of training data available in both modalities, e.g., only a few seconds of speech training data per-person and only a single training facial image per-person.

We consider fusion of the speech and facial modalities at the feature level, i.e., low level fusion. It has been argued that low level fusion allows richer information to be extracted from biometric signals than high level fusion such as at the score level [7], [8]. However this advantage must be balanced against the difficulty of feature level fusion caused by differences in feature representations between modalities [9]. Feature level fusion is often performed by concatenation of feature vectors [10], which can lead to several problems. Interpolation has been used to overcome the problem of combining features captured at different sampling rates, for example by increasing the sampling rate of video to match that of speech [11], [12]. A similar effect to interpolation can also be achieved using a coupled HMM [13] or a hybrid concatenation system [14]. Additionally, if one modality is represented by much larger feature vectors than the others, that modality will dominate in any comparison using the concatenated features. Typically, problems arising from large feature representations are mitigated by using dimensionality reduction techniques such as principal component analysis (PCA), linear discriminant analysis (LDA) [15], [10], tensor based methods [16], manifold learning techniques [17], or neural networks [18], [19]. The use of certain dimensionality reduction methods may not be possible given only a limited amount of training data. This small sample size problem has also been addressed for multimodal data, using geometry preserving projections [20].

In this paper we present a novel method for multimodal combination and comparison, which allows features of multiple modalities to be combined despite vast differences in both data rates and feature sizes. Specifically, we combine a single facial image with multiple speech frames, where the facial image is represented by very large Gabor features and each speech frame is represented using compact sub-band spectral features. This can be thought of as interpolation of the single face image over multiple speech frames. We then use modified cosine similarity to compare the bimodal feature vectors for person identification, which provides a normalized similarity for both modalities despite the imbalance in feature sizes.

We further consider scenarios in which there is limited training data in both the facial and speech

modalities. While facial identification is often performed using just a single training image per-person [21], [22], [23], [24], speaker identification is usually performed using a statistical model, such as a Gaussian mixture model (GMM) or hidden markov model (HMM), assuming there is a significant amount of training data from each speaker. In the case of limited training data, it may not be feasible to reliably estimate the parameters of such a statistical model directly. As a solution, parameters can be adapted from a universal background model (UBM) [25], [26]. Recently, speaker identification methods based on similarity rather than probability, such as fuzzy vector quantization [27], [28], have been developed as an alternative solution to the problem of limited training data. The work in [29] has discussed the selection of discriminative features from limited training data for speaker recognition. An earlier non-statistical approach used a linear classifier for separating the true speaker from imposters [30].

Finally, we consider the presence of both environmental audio noise and facial occlusion, under which conditions speaker and face recognition become much more difficult problems. With *a priori* knowledge of the noise characteristics it may be possible to remove the effect of the noise from recognition, using speech enhancement or noise compensation techniques (e.g., spectral subtraction [31] or Wiener filtering [32]). Additionally, robust features, e.g., RASTA [33], missing feature theory and multi-style training [34], [35], [36] have also been used to develop noise-robust speaker recognition systems. Most of these systems are built using statistical speaker models (e.g., GMMs or HMMs). In this paper, we extend the study of noise robustness into similarity-based speaker recognition, to tackle the problem of speaker recognition with limited training data. Specifically, we build on previous work [36] to develop a similarity-based framework which combines multicondition training and optimal feature selection. We aim to develop a method of robust speaker recognition in the presence of time-varying noise without assuming any specific information about the noise, and with limited training data.

Like noise in speech, occlusion of face images can dramatically degrade recognition accuracy. Partial occlusion is commonly dealt with by dividing the face into smaller sub-images, each of which can be recognised separately. The recognition scores from the sub-images can then be combined in a way that minimizes the contributions of the corrupted sub-images. Methods for the combination include voting systems [37], [38] and weighted averages [39]. It is also possible to use patch-based HMMs to model the face while ignoring occlusions [40]. Another recent approach based on sparse representation also shows robustness to partial occlusion but to achieve sparsity, this approach requires large number of training images per person [41]. In this paper, we study the problem of partially occluded face recognition as part of our bimodal system, within the framework of multicondition training with optimal feature selection. We aim to develop a system capable of recognizing a person with partially occluded face images without

assuming specific information about the occlusion, and with a single training image for each person.

Many of the above problems, e.g., speaker recognition using noisy data, face recognition using occluded images, lack of prior knowledge about the noise or occlusion, limited training data, and bimodal fusion with balanced contributions from modalities with different data rates and feature sizes, have been studied previously. In this paper, we consider *all* of these problems simultaneously within a single framework. The paper is organized as follows. In Section II, we introduce the new method for combining speech and facial image features, and the modified cosine similarity for person recognition with limited training data for both modalities. Section III presents the framework which incorporates multicondition training and missing feature theory into the similarity-based system, for robustness to simultaneous corruption of both modalities without assuming specific information about the corruption. Experimental studies of person identification are presented in Section IV. Finally, conclusions are drawn in Section V.

II. Similarity-Based Bimodal Person Recognition

We consider the representation of a person by using a *short* speech segment and a *single* face image from the person. Let λ be the index of a person, belonging to a group Λ of persons of interest. Each person λ has a single face image I^λ and a short speech segment of T frames $X^\lambda = (x^\lambda(1), x^\lambda(2), ..., x^\lambda(T))$ for training, where $x^\lambda(t)$ is the training speech frame at time t. To accommodate corruption in either or both of the modalities, we represent each speech frame as F non-overlapped subbands, i.e., $x^\lambda(t) = (x_1^\lambda(t), x_2^\lambda(t), ..., x_F^\lambda(t))$ where $x_f^\lambda(t)$ is the feature for subband f in frame $x^\lambda(t)$. Similarly, we represent each face image as K non-overlapped sub-images, i.e., $I^\lambda = (I_1^\lambda, I_2^\lambda, ..., I_K^\lambda)$ where I_k^λ is the feature for the kth sub-image. In this way, corrupted subbands and/or sub-images can be removed from the representation without affecting the other useful subbands and sub-images. To overcome the problem of differing data rates between the two modalities, i.e., in order to combine a single static face image with a time sequence of speech frames, we combine the face image I^λ with every speech frame $x^\lambda(t)$ to form a new bimodal time sequence $\mathbf{X}^\lambda$:

$$\begin{aligned}\mathbf{X}^\lambda &= \{(x^\lambda(1), I^\lambda), (x^\lambda(2), I^\lambda), ..., (x^\lambda(T), I^\lambda)\} \\ &= \{\mathbf{x}^\lambda(1), \mathbf{x}^\lambda(2), ..., \mathbf{x}^\lambda(T)\}\end{aligned} \tag{1}$$

In this new bimodal time sequence, each bimodal frame $\mathbf{x}^\lambda(t)$ groups together the speech subbands at time t and the whole face image represented by the sub-images:

$$\mathbf{x}^\lambda(t) = (x_1^\lambda(t), x_2^\lambda(t), ..., x_F^\lambda(t), I_1^\lambda, I_2^\lambda, ..., I_K^\lambda) \tag{2}$$

This representation allows for a single static image to be combined with an arbitrary number of speech frames, eliminating the problem of differing data rates. [1] Although in this paper we study only the combination of a single face image with many speech frames, our system is general enough to be used with video of the face. In the training stage, we create the bimodal representation $\mathbf{X}^\lambda$ for each person $\lambda \in \Lambda$.

In recognition, let $Y = (y(1), y(2), ..., y(\Gamma))$ be a test speech segment of Γ frames and J be a test image, from an unknown person. In the same way as before, we represent each speech frame $y(t)$ in subbands $y(t) = (y_1(t), y_2(t), ..., y_F(t))$ and the face image J in sub-images $J = (J_1, J_2, ..., J_K)$, to form a bimodal time sequence for the unknown person $\mathbf{Y} = \{\mathbf{y}(1), \mathbf{y}(2), ..., \mathbf{y}(\Gamma)\}$, where each bimodal frame $\mathbf{y}(t) = (y_1(t), y_2(t), ..., y_F(t), J_1, J_2, ..., J_K)$. Let $C(\mathbf{Y}, \mathbf{X}^\lambda)$ represent a similarity measure between the test sequence $\mathbf{Y}$ and a model sequence $\mathbf{X}^\lambda$ for person λ. We identify the unknown person as follows, assuming text-independent training and test speech segments

$$\begin{aligned} \hat{\lambda} &= \arg\max_{\lambda} C(\mathbf{Y}, \mathbf{X}^\lambda) \\ &= \arg\max_{\lambda} \sum_{t=1}^{\Gamma} \max_{\tau} C(\mathbf{y}(t), \mathbf{x}^\lambda(\tau)) \end{aligned} \tag{3}$$

In order to perform text-independent speaker recognition, for each test frame $\mathbf{y}(t)$ we select the best matching model frame $\mathbf{x}^\lambda(\tau)$ for comparison.

Typically, a speech frame $x^\lambda(t)$ of 20 ms long can be represented by using 30-40 features, covering 5-10 subbands (e.g., subband MFCC [36]), while a facial sub-image I_k, of 20×20 pixels for example, could be represented by more than 10^4 coefficients (e.g., Gabor features [42]). Without proper normalization, such a huge disparity in feature sizes may cause the features from one modality to completely dominate in the comparison. In the following, we introduce a novel similarity measure, modified cosine similarity, for combining and comparing modalities of different sizes which effectively overcomes this problem.

The cosine similarity $C(\mathbf{a}, \mathbf{b})$ between two vectors $\mathbf{a}$ and $\mathbf{b}$ can be expressed as

$$C(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||||\mathbf{b}||} \tag{4}$$

If $\mathbf{a}$ and $\mathbf{b}$ are each divided into Q local vectors $\mathbf{a} = (a_1, a_2, ..., a_Q)$ and $\mathbf{b} = (b_1, b_2, ..., b_Q)$, (for example, if $\mathbf{a}$ and $\mathbf{b}$ each represent an image which is divided into Q smaller sub-images), the cosine similarity between $\mathbf{a}$ and $\mathbf{b}$ can be expressed in terms of the cosine similarities between the Q corresponding local

[1] This representation can also be thought of as interpolation of a single face image over a number of speech frames.

vectors, i.e.,

$$\begin{aligned} C(\mathbf{a},\mathbf{b}) &= \sum_{q=1}^{Q} \frac{a_q \cdot b_q}{||\mathbf{a}||||\mathbf{b}||} \\ &= \sum_{q=1}^{Q} \frac{a_q \cdot b_q}{||a_q||||b_q||} \frac{||a_q||||b_q||}{||\mathbf{a}||||\mathbf{b}||} \\ &= \sum_{q=1}^{Q} \frac{a_q \cdot b_q}{||a_q||||b_q||} w_q \\ &= \sum_{q=1}^{Q} C(a_q,b_q) w_q \end{aligned} \tag{5}$$

where $C(a_q,b_q) = a_q \cdot b_q / ||a_q||||b_q||$ is the cosine similarity between local vectors a_q and b_q. From (5) we can see that the overall cosine similarity is the sum of all the local cosine similarities $C(a_q,b_q)$ weighted by w_q, which equals the norms of the appropriate local vectors compared to the norms of the overall vectors. As the weight w_q is a function of the overall norms, it will be affected by any local vector corruption in either $\mathbf{a}$ or $\mathbf{b}$. In other words, the weighting can spread local vector corruptions globally. To avoid this problem, we assume a uniform weight w_q for all the local vectors, meaning they contribute equally to the overall similarity. Thus, we use a uniformly-weighted cosine similarity to compare the two multimodal frames, $\mathbf{y}(t)$ and $\mathbf{x}^{\lambda}(\tau)$, required in (3). This can be written as

$$C(\mathbf{y}(t),\mathbf{x}^{\lambda}(\tau)) \simeq \sum_{f=1}^{F} C(y_f(t),x_f^{\lambda}(\tau)) + \sum_{k=1}^{K} C(J_k,I_k^{\lambda}) \tag{6}$$

Note that, since each similarity measure $C(y_f(t),x_f^{\lambda}(\tau))$ of speech subbands and $C(J_k,I_k^{\lambda})$ of facial sub-images in (6) varies in the same range, from -1 to 1, all speech subbands and face sub-images contribute equally to the overall similarity, independent of the vast size disparity between the two local modality features.

Equation (6) can be expressed in an equivalent form

$$\begin{aligned} p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau)) &= H^{C(\mathbf{y}(t),\mathbf{x}^{\lambda}(\tau))} \\ &= \prod_{f=1}^{F} H^{C(y_f(t),x_f^{\lambda}(\tau))} \prod_{k=1}^{K} H^{C(J_k,I_k^{\lambda})} \end{aligned} \tag{7}$$

where $H > 1$ is a positive base number. The function $p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau))$ shares the characteristics of an exponent-type likelihood function for the test frame $\mathbf{y}(t)$ associated with person λ, represented by the model frame $\mathbf{x}^{\lambda}(\tau)$. Correspondingly, the recognition decision rule (3) can be written in an equivalent

form

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{t=1}^{\Gamma} \max_{\tau} \log p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau)) \tag{8}$$

We will use this equivalent form to introduce robustness into the recognition system, against corruption in either or both of the modalities. At this point, the new recognition system (8) is capable of accommodating speech and image features of different data rates, different feature sizes (e.g., long Gabor feature vectors for facial images versus short spectral feature vectors for speech frames), and accommodating limited training examples for both modalities (e.g., a short training speech segment, and as few as a single training image, for each person). In all our experiments, H is defined as 1.5×10^4.

III. Robustness To Corruption

The system as currently defined assumes that both the training and test data of the speech and face image are uncorrupted. We extend the system to be resistant both to background noise for the speech modality and to partial occlusion for the facial modality. We achieve this by firstly modifying the computation of the likelihood $p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau))$ of a noisy bimodal test frame $\mathbf{y}(t)$ associated with a clean bimodal model frame $\mathbf{x}^{\lambda}(\tau)$, to incorporate multicondition training. Let $\mathbf{X}^{\lambda} = (\mathbf{x}^{\lambda}(1), \mathbf{x}^{\lambda}(2), .., \mathbf{x}^{\lambda}(T))$ be the given clean bimodal training sequence for person λ, and $\mathbf{X}^{\lambda,i} = (\mathbf{x}^{\lambda,i}(1), \mathbf{x}^{\lambda,i}(2), .., \mathbf{x}^{\lambda,i}(T))$, $i = 0, 1, ..., L$, represent $L + 1$ multicondition training sequences generated from $\mathbf{X}^{\lambda}$, where each $\mathbf{X}^{\lambda,i}$ simulates a different corruption condition, with $\mathbf{X}^{\lambda,0} = \mathbf{X}^{\lambda}$ corresponding to the clean condition. These multicondition training sequences are combined to model a test bimodal sequence $\mathbf{Y}$ with feature corruption. The likelihood of a noisy test frame $\mathbf{y}(t)$ associated with a clean model frame $\mathbf{x}^{\lambda}(\tau)$ given multicondition training can be written as

$$p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau)) = \sum_{i=0}^{L} p(\mathbf{y}(t)|\mathbf{x}^{\lambda,i}(\tau))P(i|\lambda) \tag{9}$$

where $p(\mathbf{y}(t)|\mathbf{x}^{\lambda,i}(\tau))$ is the likelihood of the noisy test frame $\mathbf{y}(t)$ associated with the model frame $\mathbf{x}^{\lambda(\tau),i}$ corrupted at condition i, and $P(i|\lambda)$ represents our prior knowledge (e.g., a prior probability) of the corruption condition for person λ, which we assume to be a uniform distribution across all the training conditions. In our experiments, instead of assuming *a priori* knowledge about the test data corruption, for both speech and face, we try to compensate for a wide range of corruptions by properly generating the multicondition training data $\mathbf{X}^{\lambda,i}$ and performing optimal feature selection in the recognition. For example, we add wide-band noise to the clean speech training data at different signal-to-noise ratios (SNRs) to simulate a broad range of acoustic noises. This scheme is combined with optimal feature

selection, described below, in the decoding stage to make the system robust to corruption conditions unseen in the training stage.

Given the test data, in order to reduce mismatches between the simulated corruption and actual corruption, we introduce optimal feature selection. Instead of comparing the full sets of model and test features to calculate the frame likelihood (i.e., (9)), which is subject to the contamination of training and test condition mismatch, we will determine the frame likelihood based only on the 'matched' local features in terms of large likelihoods (these features are likely to correspond to those with matched training and testing conditions). Then, we perform recognition in favor of the person who has the maximum number of matched local features under the same multi-style training conditions for all the persons. Our algorithm can be viewed as an implementation of missing-feature theory or the recognition-by-parts principle [36] [43] without assuming knowledge of the identities of the matched-condition features.

Consider the test frame $\mathbf{y}(t)$ and model frame $\mathbf{x}^{\lambda,i}(\tau)$ under corruption condition i. The matched-feature likelihood can be expressed as $p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau))$, where $\mathbf{y_i}(t) \subseteq \mathbf{y}(t)$ is a test feature subset containing speech subbands and facial sub-images which match the corresponding model feature subset $\mathbf{x_i}^{\lambda,i}(\tau) \subseteq \mathbf{x}^{\lambda,i}(\tau)$, where $\mathbf{i} = \{f\} \cup \{k\}$ is the index set defining the local speech and image features in the two matched sets, with $f \in (1, 2, ..., F)$ and $k \in (1, 2, ..., K)$. We can obtain an estimate of $\mathbf{x_i}^{\lambda,i}(\tau)$, and hence $\mathbf{y_i}(t)$, for each person λ at each simulated corruption condition i, by maximizing a "normalized likelihood" function $P(\mathbf{x_i}^{\lambda,i}(\tau)|\mathbf{y_i}(t))$ defined in (10) below. Assuming an equal prior probability P for all model feature subsets, this normalized likelihood function is defined as

$$\begin{aligned} P(\mathbf{x_i}^{\lambda,i}(\tau)|\mathbf{y_i}(t)) &= \frac{p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau))P}{p(\mathbf{y_i}(t))} \\ &= \frac{p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau))P}{\sum_{\mathbf{x_i}\in\text{training}} p(\mathbf{y_i}(t)|\mathbf{x_i})P + \sum_{\mathbf{x_i}\notin\text{training}} p(\mathbf{y_i}(t)|\mathbf{x_i})P} \\ &\simeq \frac{p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau))}{\sum_{\lambda'\in\Lambda}\sum_{\tau'=1}^{T'}\sum_{i'=0}^{L} p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda',i'}(\tau')) + \epsilon} \end{aligned} \tag{10}$$

In the denominator, the first term is the average likelihood of the test feature subset, counting all the training persons, training frames and training corruption conditions (including the clean training condition); the second term, ϵ, accounts for the likelihood of any test feature subset without matching training examples (hence both the numerator and the first term of the denominator approach zero). [2]

[2]We find that with multicondition training for all the persons (i.e., the first term), the likelihood ϵ can become very small and hence insignificant in the recognition. Therefore, for simplicity, in the following derivation we assume $\epsilon = 0$. However, in our experiments we set ϵ to a small value 10^{-5} to prevent any possibility of numerical division by zero.

Note that if each $p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau))$ is a probabilistic likelihood, then (10) takes the form of Bayes' rule. We can show that maximizing the above defined normalized likelihood leads to an optimal estimate of the matched feature subset under each training condition, in the sense that larger normalized likelihoods are obtained when more correctly matched features are found. To show this, assume $\mathbf{y_i}(t)$ and $\mathbf{x_i}^{\lambda,i}(\tau)$ to be two matched feature sets in terms of the likelihoods $p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau)) \geq p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda',i'}(\tau'))$ for any $\mathbf{x_i}^{\lambda',i'}(\tau') \neq \mathbf{x_i}^{\lambda,i}(\tau)$. Express $\mathbf{y_i}(t)$ as a union of any subset $\mathbf{y_{i_1}}(t) \subset \mathbf{y_i}(t)$ and the complement $\mathbf{y_{i_2}}(t)$, and $\mathbf{x_i}^{\lambda,i}(\tau)$ as a union of the corresponding matching subsets $\mathbf{x_{i_1}}^{\lambda,i}(\tau)$ and $\mathbf{x_{i_2}}^{\lambda,i}(\tau)$. We have, for any $\mathbf{x_i}^{\lambda',i'}(\tau') \neq \mathbf{x_i}^{\lambda,i}(\tau)$:

$$\begin{aligned}\frac{p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau))}{p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda',i'}(\tau'))} &= \frac{p(\mathbf{y_{i_1}}(t)|\mathbf{x_{i_1}}^{\lambda,i}(\tau))p(\mathbf{y_{i_2}}(t)|\mathbf{x_{i_2}}^{\lambda,i}(\tau))}{p(\mathbf{y_{i_1}}(t)|\mathbf{x_{i_1}}^{\lambda',i'}(\tau'))p(\mathbf{y_{i_2}}(t)|\mathbf{x_{i_2}}^{\lambda',i'}(\tau'))} \\ &\geq \frac{p(\mathbf{y_{i_1}}(t)|\mathbf{x_{i_1}}^{\lambda,i}(\tau)}{p(\mathbf{y_{i_1}}(t)|\mathbf{x_{i_1}}^{\lambda',i'}(\tau'))} \end{aligned} \tag{11}$$

The last inequality is obtained because $p(\mathbf{y_{i_2}}(t)|\mathbf{x_{i_2}}^{\lambda,i}(\tau)) \geq p(\mathbf{y_{i_2}}(t)|\mathbf{x_{i_2}}^{\lambda',i'}(\tau'))$ based on the above assumption for the matched feature sets. Rewriting the normalized likelihood (10) in terms of the likelihood ratios, we obtain

$$P(\mathbf{x_i}^{\lambda,i}(\tau)|\mathbf{y_i}(t)) \simeq \frac{1}{\sum_{\lambda'\in\Lambda}\sum_{\tau'=1}^{T'}\sum_{i'=0}^{L}\frac{p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda',i'}(\tau'))}{p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau))}} \tag{12}$$

Applying the likelihood ratio inequality (11) to this expression, we can obtain an inequality concerning the normalized likelihoods of the matched feature sets with different sizes

$$P(\mathbf{x_{i_1}}^{\lambda,i}(\tau)|\mathbf{y_{i_1}}(t)) \leq P(\mathbf{x_i}^{\lambda,i}(\tau)|\mathbf{y_i}(t)) \tag{13}$$

This inequality indicates that larger normalized likelihoods are obtained when more features are matched.

Note that the likelihood $p(\mathbf{y_i}(t)|\mathbf{x_i}^{\lambda,i}(\tau))$ and the normalized version $P(\mathbf{x_i}^{\lambda,i}(\tau)|\mathbf{y_i}(t))$ are proportional to each other. So substituting the matched-feature normalized likelihood $P(\mathbf{x_i}^{\lambda,i}(\tau)|\mathbf{y_i}(t))$ into (9) to replace the full feature set likelihood, and further substituting the resulting multicondition frame likelihood into (8) to replace the single training condition frame likelihood, we can obtain a new recognition rule which seeks to find the most-likely person by jointly maximizing the normalized likelihood over all persons and all possible feature subsets for all the test frames:

$$\hat{\lambda} = \arg\max_{\lambda}\sum_{t=1}^{\Gamma}\max_{\tau}\log\Big\{\sum_{i=0}^{L}\max_{\mathbf{i}}P(\mathbf{x_i}^{\lambda,i}(\tau)|\mathbf{y_i}(t))\Big\} \tag{14}$$

As understood, this new recognition system (14) favors the person with the maximum number of matched features given the same training conditions for all the persons.

We say (14) is a recognition system which combines both multicondition training and optimal feature estimation, with the aim of offering robustness to unknown test corruption in either or both of the modalities. Ideally, multicondition training will cover the expected range of corruption, and optimal feature selection will refine the modeling accuracy for each specific corruption, by focusing on the local features likely to have matched corruption conditions. The system selects the person with the maximum number of matched features to minimize the error of recognition. Through optimal feature selection, the system is effectively changing the weighting given to each modality, to reflect the amount of reliable information that can be extracted from each modality at every frame.

The computational complexity of the above system will grow linearly in the number of simulated noise conditions (i.e., L). However, the use of optimal feature selection together with multicondition training allows for a very large number of realistic noise conditions to be accommodated given only a small number of simulated noise conditions, meaning a reduced computational cost. Additionally, the feature selection in each frame can be computed efficiently (see a demonstration in [44] for face recognition based on GMM face models) and independently, meaning the robust multimodal recognition system can be parallelised.

IV. Experimental Studies

In this section we evaluate the performance of the proposed system for bimodal person identification with limited training data and corruption of both modalities. Two databases were used in these experiments. First, the SPIDRE speaker database [45], which is a subset of the Switchboard database and consists of four telephony conversation halves from 45 speakers (27 male, 18 female), was used to provide the speech modality. Second, the AR facial image database [46], which consists of the frontal face images of 126 persons with realistic partial occlusions, was used to provide the facial modality. Note in many of the experiments discussed later, we used only a single short segment of speech or a single frame of facial image to represent a person. Therefore the phrase *training a model* for each person is effectively a process of indexing the person to their available sample data.

Our experiments include three major parts. First, we considered the speech modality alone and compared the proposed new system, the modified cosine-similarity based system, with conventional systems for speaker recognition. This comparison is focused on robustness both to limited training data and to variable types of noise corruption. Second, we considered the image modality alone and compared the new system with previously published systems for face recognition. This comparison is focused on robustness to realistic partial occlusion given limited training data. Finally, we conducted bimodal

recognition experiments where either a single or both modalities were corrupted to varying degrees, given limited training data for both modalities. These last tests show how using both modalities together can improve recognition accuracy over the unimodal systems in adverse conditions.

A. Speaker Identification with Noise Corruption

In this study, we focused on the noise problem using the SPIDRE database. As in [45] and [47], we used the two conversation halves from the same handset, one for training and the other for testing. Each speech sample was divided into 20 ms frames overlapping by 10 ms. Each frame, of a bandwidth of 4 kHz, was processed through a 22-channel log mel-scale filterbank and the filter outputs decorrelated with a high-pass filter, giving 21 decorrelated log mel filterbank coefficients. These coefficients were uniformly placed into groups of three, giving seven subband features. First-order derivative coefficients were included, resulting in 14-subband feature streams for each frame, each stream containing three elements. That is, each speech frame $x^{\lambda}(t) = (x_1^{\lambda}(t), x_2^{\lambda}(t), ..., x_F^{\lambda}(t))$ for speaker λ is represented by $F = 14$ subband feature streams corresponding to seven subbands; each subband feature stream $x_f^{\lambda}(t)$ contains three elements, corresponding to the static or first-order derivative of a subband. For the new system, speaker models were constructed using segments of speech from each speaker of varying durations from 1 s to 30 s, and speaker identification experiments were performed using three test speech samples of durations varying from 1 s to 30 s from each speaker. We compared the performance of the new system, given limited training data, against previously published results based on a GMM system. The GMM system contained 32 Gaussian densities with diagonal covariance matrices for each speaker, and was trained using all the available training data (about two minutes) for each speaker [47].

First, we compared speaker identification accuracy using the clean training and testing data. Table I presents the results for the proposed new system, as a function of both the training data duration and test data duration. The results are compared to that of the GMM system, cited from [47]. From these results we can see that the identification accuracy of our new system with 30 s of training data is comparable with the GMM-based result. In fact when the new system is trained with 30 s of speech and tested with 10 s of speech its identification accuracy exceeded that of the GMM system with about 2 min of training data. The results indicate that the new recognition system based on modified cosine similarity is a viable method for speaker identification.

By examining the information from Table I in graphical form, shown in Fig. 1, we can make several observations about the relationship between identification accuracy and testing/training duration. Given a short testing duration, e.g., 1 s of test data, reasonable identification accuracy can be achieved if there is

sufficient training (Fig. 1(a)). If however the training duration is very short, increasing testing duration does not necessarily improve accuracy (Fig. 1(b)). Examining Fig. 1(b) we see that increasing training length quickly increases accuracy for a given testing duration. However, in Fig. 1(a) we see that increasing testing duration yields improvement only when there is sufficient training data available, for example, with only 1 s or 2 s of training, increasing testing duration does not significantly improve accuracy.

Next, we assess the effectiveness of optimal feature selection for noise robust speaker recognition. We considered band-limited corruption and assumed no knowledge of the noise band location. The new system used the maximum-probability algorithm, shown in (14), to select the reliable subband features from the corrupted test samples during recognition. In this experiment, we used only clean training data to construct the speaker models (i.e., $L = 0$ in (14)). We used an oracle model to help assess the effectiveness of the optimal feature selection algorithm. The oracle model used prior knowledge of the noise band location to remove the corrupted subbands before performing recognition. In addition, a baseline system which performed recognition using all the subbands regardless of their level of corruption was also tested; this is referred to as the 'do-nothing' system.

Noise corrupted test samples were generated by adding narrow-band noise to clean speech test samples at an SNR of 0 dB. Different narrow-band noises, of different central frequencies and bandwidths, were created to corrupt different subbands and different numbers of adjacent subbands. Speaker models were constructed using 30 s of clean speech from each speaker, and speaker identification experiments were performed using five 10 s samples from each speaker. The results of this experiment are shown in Table II, for variable narrow-band noises at different frequency locations and affecting different numbers of subbands, within the seven subbands of the representation.

Table II shows that in most of the cases the new system performed better than the oracle model and always significantly better than the 'do-nothing' model. This indicates that the system is capable of removing the contribution of noise corrupted speech subbands from each speaker's score. The fact that the new system could outperform the oracle model in many cases, may be due to the fact the oracle model completely removes all bands believed to be corrupted by noise. It may be the case that some subband features, for example, located at the edges of the noise bands or having a high local SNR, are only partially corrupted and thus are still usable for recognition. The new system performing automatic feature selection to maximize probability may be taking advantage of this additional information, ignored by the oracle model, to produce more accurate identification scores.

Finally, we evaluate the performance of the new system combining both multicondition training and optimal feature selection, i.e., (14) with $L > 0$, for dealing with more difficult, full-band, nonstationary

noise corruption without assuming prior knowledge of the characteristics of the noise. For each speaker, a multicondition model was constructed by adding low-pass filtered white noise, with a 3 dB cut-off frequency of 2 kHz, to each clean training segment at SNRs of 10 dB, 15 dB and 20 dB, respectively. Thus, with the clean training condition included, we have a total of four (i.e., $L = 3$) training conditions in the system (14). Two training segments, of duration of 5 s and 10 s respectively, were used to construct the multicondition model for each speaker.

Noise corrupted test samples were created by adding real-world non-stationary full-band noise to clean speech test samples at SNRs of 10 dB, 15 dB and 20 dB, respectively. Three noise types were used, which were restaurant, street and pop-song. Fig. 2 shows the noise spectrograms, indicating the wide-band, time-varying natures of these noises. Ten testing samples, with durations of 5 s and 10 s respectively, were generated for each speaker. In addition to tests of the new system, tests were performed with the system using multicondition training only (i.e., (14) with $L = 3$ but without the $\mathbf{i}$ selection), optimal feature selection only (i.e., (14) with the $\mathbf{i}$ selection but with $L = 0$), and a baseline 'do-nothing' model.

The results are presented in Table III. We can see that the new system significantly outperformed the do-nothing model at all the low SNR conditions. There are only a few exceptions for the pop-song noise, at the higher SNR of 20 dB, where no improvement or only small improvement was found for the new system in comparison to the do-nothing model, indicating the difficulty of modeling speech-like noise for accurate speaker recognition. The new system also suffered some performance loss at the clean testing condition, as typically experienced by most noise-robust systems. We also see that combining multicondition training and optimal feature selection in the new system led to greater improvement in system performance than provided by either of the techniques in isolated operation, compared to the do-nothing model. This is especially true for the low SNR tests, and for the tests with shorter training speech segments.

B. Face Identification with Realistic Partial Occlusion

The AR facial database contains realistic partial occlusions by sunglasses and scarf. From the database, we randomly selected 45 persons (the same number of persons as in the SPIDRE speech database) for testing. Each face image, of 165×120 pixels, was processed by a Gabor filterbank at four scales and four orientations. The Gabor filterbank outputs were down-sampled by two times and each filterbank output was split into 16 equal sized blocks. Corresponding blocks from each filterbank output were concatenated so that the face was represented by 16 feature vectors each 2400 elements long. That is, each face image $I^\lambda = (I_1^\lambda, I_2^\lambda, ..., I_K^\lambda)$ for person λ is split into $K = 16$ sub-images, and each sub-image I_k^λ is represented

using 2400 feature elements (thus, each face image was represented by a total of $2400 \times 16 = 38400$ elements). Note the large difference in feature size between a speech subband $x_f^\lambda(t)$ (three elements) and a facial sub-image I_k^λ (2400 elements). In this experiment, a single clean face image, randomly selected from the training set, was used as the training image for each person (i.e., $L = 0$ in (14)). Tests were carried out using three clean face images and three occluded face images from each occlusion condition – sunglasses and scarf – for each person.

The results of the new recognition system are shown in Table IV. These results are comparable to or exceeded those in the recent literature for different recognition systems [24], [23], [40], [22]. Results of these previous systems are included in Table IV for comparison. These previous systems were trained on the same database using one image per-person and tested using images with the same occlusions.

C. Bimodal Person Identification with Unknown Corruption of Both Modalities

Finally, bimodal recognition experiments were conducted assuming limited training data for both modalities, and corruption of either a single or both modalities by realistic noise to varying degrees. The full version of the new system (14) was used, in which each frame $\mathbf{x}^\lambda(t)$ for person λ combines speech subband features and facial sub-images features, as defined in (2). More specifically, each frame contained $F = 14$ subband feature streams and $K = 16$ sub-image features, in the form of decorrelated log magnitude spectra for speech subbands and Gabor coefficients for sub-images, as detailed above. In the experiments, each person's face model consisted of a single clean face image, the same as in Section IV-B; this was then combined with the person's speech model built using multicondition training speech data with simulated noise, the same as in Section IV-A, with four corruption conditions. Thus, the new system, is expected to be able to deal with both full-band noise corruption in the speech modality (by combining multicondition training and optimal subband selection), and partial occlusion in the facial modality (by optimal sub-image selection), without requiring prior knowledge about the corruptions. Testing was performed for each person by taking three speech samples of either 5 s or 10 s from each noise condition (restaurant, street, pop song), paired with a randomly chosen face image from one of the three facial occlusion conditions (clean, sunglasses, and scarf). The new system was compared to a 'do-nothing' model, which performed recognition using all the feature components from both modalities, without applying noise compensation. An example of the bimodal test data used in this experiment is shown in Fig. 3.

Firstly, recognition experiments were conducted assuming corruption of only a single modality of the two modalities. The recognition results using occluded test images and clean test speech are shown in

Table V, as a function of the occlusion type (including no occlusion) and the training and testing speech duration. Compared to the face-only recognition results shown in Table IV, we see that inclusion of the clean speech samples from the subjects improves the recognition accuracy for both occlusion types. The improvement is observed for both the new system and the baseline do-nothing system, and observed for the variable training and testing speech durations examined in the experiments. The new system further improves over the do-nothing system. For the new system, the recognition accuracy with 10 s training speech is greater than or equal to the accuracy obtained with 5 s training speech. However, in common with the speaker recognition experiments, we do not observe a strong correlation between increased testing speech duration and increased recognition accuracy. Next, recognition experiments were performed using noisy test speech and clean test images, as a function of the noise type and SNR, and the durations of training and testing speech (the same conditions as shown in Table III). An accuracy rate of 100% was obtained for all these conditions for both the new system and the baseline do-nothing system.

Then, recognition experiments were conducted assuming corruption on both modalities. The results are summarized in Table VI, as a function of the training and testing speech durations, the acoustic noise, the SNR, and the facial occlusion. In Table VI, we see that combining the speech or facial information with the other modality helped improve recognition accuracy even if the combined information was corrupted. When both modalities are corrupted, the recognition accuracy of the bimodal system remains higher than the best accuracy achieved by either unimodal system tested on corresponding corruption conditions. For example, in the case of restaurant noise with SNR = 10 dB, with 10 s speech for training and 10 s speech for testing, and with face being occluded by sunglasses, the unimodal speaker and facial identification systems scored 63.7% and 89.6% respectively (see Table III and Table IV), while the bimodal system scored 100% given identical corruption conditions. The new system combines multicondition training and optimal feature selection to give an additional improvement in accuracy compared to straightforward multimodal combination. This additional accuracy improvement can be seen by comparing the results in the ‘do nothing’ columns with the results produced by the new system. For example, with face occluded by sunglasses and SNR = 10 dB street noise, with 5 s training and 5 s testing, the ‘do-nothing’ system scored 97.0%, while under the same conditions the new system scored 100% accuracy. By being able to extract the maximum number of matched features from each modality, the new system is dynamically changing, at each frame, the weighting given to each modality depending on how much reliable information is present. A similar improvement in the accuracy of audio-visual person identification by dynamically weighting each modality, depending on reliability, has been observed in [8]. As with the previous experiments, we

see that longer speech training often produces greater recognition accuracy than shorter speech training. The new system outperformed the 'do-nothing' model in all the test conditions. These results demonstrate that the new system is capable of improving recognition accuracy compared to the component unimodal systems, as well as improving accuracy compared to the bimodal systems based on multicondition training or missing-feature decoding alone.

V. Conclusions

Person identification becomes a challenging problem when there is limited training data available for the person, and when the data used for identification is corrupted, as will happen in many realistic applications. By using the information from multiple modalities together it is possible to increase robustness to data sparsity and corruption, and hence increase identification accuracy. In this paper we have proposed a new method of bimodal person identification that can be used with limited training data in both modalities, on the order of a few seconds of speech and a single facial image per person, and that is robust to realistic corruption of both modalities. The new method combines the speech and face image at the feature level, and compares the bimodal features using a modified cosine similarity. The new similarity measure balances the contributions of the two modalities despite their vast differences in both data rates and feature sizes, and offers the capability of accommodating very limited training data for both modalities. Robustness to corruption in either or both modalities is obtained by incorporating multicondition model training and optimal feature selection. The system is designed to provide corruption robustness assuming minimum prior information about the specific corruption. Experiments were performed on two challenging databases, the AR face database and SPIDRE speaker database, with various types of realistic partial facial occlusion and nonstationary acoustic noise without assuming prior information. The experimental results demonstrate that the new method of bimodal combination offered improved person identification accuracy compared to other systems. Future work will focus on the extension of multicondition training to the facial modality for dealing with lighting and/or pose variations.

REFERENCES

[1] K. W. Bowyer, K. I. Chang, P. Yan, P. J. Flynn, E. Hansley, and S. Sarkar, "Multi-modal biometrics: An overview," *Second Workshop on Multi-Modal User Authentication*, 2006.

[2] P. Kartik, R. Vara Prasad, and S. Mahadeva Prasanna, "Noise robust multimodal biometric person authentication system using face, speech and signature features," *Annual IEEE India Conference*, pp. 23–27, Dec. 2008.

[3] N. Poh and J. Kittler, "A unified framework for multimodal biometric fusion incorporating quality measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011 (Preprint).

[4] M. Bendris, D. Charlet, and G. Chollet, "Introduction of quality measures in audio-visual identity verification," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1913–1916, Apr. 2009.

[5] N. Poh and J. Kittler, "A family of methods for quality-based multimodal biometric fusion using generative classifiers," *International Conference on Control, Automation, Robotics and Vision*, pp. 1162–1167, 2008.

[6] H. Choi and M. Shin, "Learning radial basis function model with matching score quality for person authentication in multimodal biometrics," *First Asian Conference on Intelligent Information and Database Systems*, pp. 346–350, Apr. 2009.

[7] D. L. Hall, *Mathematical Techniques in Multisensor Data Fusion*. Norwood, MA, USA: Artech House, Inc., 1992.

[8] N. Fox, R. Gross, J. Cohn, and R. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *IEEE Transactions on Multimedia*, vol. 9, pp. 701–714, Jun. 2007.

[9] A. Ross and A. K. Jain, "Multimodal biometrics: An overview," *Communications of The ACM*, 2004.

[10] C. Chibelushi, J. Mason, and F. Deravi, "Feature-level data fusion for bimodal person recognition," *International Conference on Image Processing and Its Applications*, pp. 399–403, Jul. 1997.

[11] C. Bregler, S. Omohundro, and Y. Konig, "A hybrid approach to bimodal speech recognition," *Conference Record of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers*, pp. 556–560, Oct 1994.

[12] G. Potamianos and H. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 3733–3736, May 1998.

[13] T. Fu, X. X. Liu, L. H. Liang, X. Pi, and A. Nefian, "Audio-visual speaker identification using coupled hidden markov models," *International Conference on Image Processing*, pp. 29–32, Sept. 2003.

[14] D. Shah, K. Han, and S. Narayanan, "A low-complexity dynamic face-voice feature fusion approach to multimodal person recognition," *IEEE International Symposium on Multimedia*, pp. 24–31, Dec. 2009.

[15] A. Ross and R. Govindarajan, "Feature level fusion using hand and face biometrics," *Proceedings of SPIE Conference on Biometric Technology for Human Identification II*, pp. 196–204, 2005.

[16] Y. Pang, X. Li, and Y. Yuan, "Robust tensor analysis with l1-norm," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 172–178, Feb. 2010.

[17] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[18] U. Meier, W. Hurst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 833–836, May 1996.

[19] P. Cosi, E. Magno Caldognetto, K. Vagges, G. Mian, and M. Contolini, "Bimodal recognition experiments with recurrent neural networks," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 553–556, Apr. 1994.

[20] T. Zhang, X. Li, D. Tao, and J. Yang, "Multimodal biometrics using geometry preserving projections," *Pattern Recognition*, vol. 41, pp. 805–813, Mar. 2008.

[21] A. James and S. Dimitrijev, "Face recognition using local binary decisions," *IEEE Signal Processing Letters*, vol. 15, pp. 821–824, 2008.

[22] Z. Li, J. Imai, and M. Kaneko, "Robust face recognition using block-based bag of words," *International Conference on Pattern Recognition*, pp. 1285–1288, Aug. 2010.

[23] R. Akbari, M. Bahaghighat, and J. Mohammadi, "Legendre moments for face identification based on single image per person," *International Conference on Signal Processing Systems*, pp. 248–252, Jul. 2010.

[24] J. Lin, J. Ming, and D. Crookes, "Robust face recognition with partially occluded images based on a single or a small number of training samples," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 881–884, Apr. 2009.

[25] P. Angkititrakul and J. H. L. Hansen, "Discriminative in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 498–508, Feb. 2007.

[26] V. Prakash and J. Hansen, "In-set/out-of-set speaker recognition under sparse enrollment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2044–2052, Sept. 2007.

[27] H. Jayanna and S. Prasanna, "Fuzzy vector quantization for speaker recognition under limited data conditions," *IEEE Region 10 Conference*, pp. 1–4, Nov. 2008.

[28] L. Lin, C. Jian, and S. Xiaoying, "A discriminative method for speaker identification with limited data," *International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 512–515, Aug. 2010.

[29] S. Kwon and S. Narayanan, "Robust speaker identification based on selective use of feature vectors," *Pattern Recognition Letters*, vol. 28, pp. 85–89, Jan. 2007.

[30] Q. Li, S. Parthasarathy, A. Rosenberg, and D. Tufts, "Normalized discriminant analysis with application to a hybrid speaker-verification system," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 681–684, May 1996.

[31] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[32] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 137–145, Apr. 1980.

[33] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[34] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 121–124, May 1998.

[35] L. Besacier, J. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, pp. 89–106, 2000.

[36] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1711–1723, Jul. 2007.

[37] C.-Y. Huang, O. Camps, and T. Kanungo, "Object recognition using appearance-based parts and relations," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 877–883, Jun. 1997.

[38] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 696–710, Jul. 1997.

[39] A. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 748–763, Jun. 2002.

[40] N.-S. Vu and A. Caplier, "Patch-based similarity HMMs for face recognition with a single reference image," *International Conference on Pattern Recognition*, pp. 1204–1207, Aug. 2010.

[41] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210–227, Feb. 2009.

[42] Y. Pang, Y. Yuan, and X. Li, "Gabor-based region covariance matrices for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, pp. 989–993, Jul. 2008.

[43] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, pp. 101–116, Sept. 2005.

[44] J. Lin, J. Ming, and D. Crookes, "Robust face recognition using posterior union model based neural networks," *IET Computer Vision*, vol. 3, pp. 130–142, Sept. 2009.

[45] D. Reynolds, "The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 113–116, May 1996.

[46] A. Martinez and R. Benavente, "The AR face database," *CVC Technical Report 24*, Jun. 1998.

[47] J. Ming, D. Stewart, and S. Vaseghi, "Speaker identification in unknown noisy conditions - a universal compensation approach," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 617–620, 2005.

PLACE PHOTO HERE

Niall McLaughlin received an M.Eng in computer science in 2008, and a Ph.D in 2012, both from Queen's University Belfast. He is currently working at Queen's University Belfast as a research associate at the Centre for Secure Information Technologies. His current research interests include robust visual tracking of persons in CCTV systems, and video analysis.

PLACE PHOTO HERE

Ji Ming (M'97) received the B.Sc. degree from Sichuan University, Chengdu, China, in 1982, the M.Phil. degree from Changsha Institute of Technology, Changsha, China, in 1985, and the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 1988, all in electronic engineering.

He was Associate Professor with the Department of Electronic Engineering, Changsha Institute of Technology, from 1990 to 1993. Since 1993, he has been with the Queen's University Belfast, Belfast, U.K., where he is currently a Professor in the School of Electronics, Electrical Engineering and Computer Science. From 2005 to 2006, he was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge. His research interests include speech and language processing, image processing, and pattern recognition.

PLACE
PHOTO
HERE

Danny Crookes was appointed to the Chair of Computer Engineering in 1993 at Queen's University Belfast, Belfast, U.K., and was Head of Computer Science from 1993-2002. He is currently Director of Research for Speech, Image and Vision Systems at the Institute of Electronics, Communications and Information Technology, Queen's University Belfast. His current research interests include the use of novel architectures (especially GPUs) for high performance speech and image processing. Professor Crookes is currently involved in projects in automatic shoeprint recognition, speech separation and enhancement, and processing of 4D confocal microscopy imagery. Professor Crookes has some 200 scientific papers in journals and international conferences.

TABLE I

SPEAKER IDENTIFICATION ACCURACY (%) ON THE SPIDRE DATABASE, COMPARING THE NEW SYSTEM AGAINST A GMM SYSTEM, AS A FUNCTION OF THE DURATIONS OF TRAINING AND TEST DATA. THE DATA FROM THIS TABLE IS DISPLAYED IN GRAPHICAL FORM IN FIG. 1.

		New system Training (s)						GMM Training (min)
		30	20	10	5	2	1	∼2
Testing (s)	30	93.3	88.9	84.4	77.8	64.4	37.8	-
	20	92.6	85.9	83.7	77.0	61.5	43.0	-
	15	93.3	86.7	82.2	76.3	63.7	39.3	91.1
	10	91.6	85.9	81.5	79.3	63.0	42.2	88.9
	5	85.9	83.7	80.7	71.9	56.3	40.0	86.7
	2	75.6	77.0	73.3	65.9	57.8	37.0	-
	1	71.9	71.1	66.7	59.3	51.1	36.3	-

TABLE II

SPEAKER IDENTIFICATION ACCURACY (%) ON THE SPIDRE DATABASE, WITH NARROW-BAND NOISE CORRUPTION AFFECTING DIFFERENT SUBBANDS AND DIFFERENT NUMBERS OF SUBBANDS, COMPARING THE NEW SYSTEM AGAINST AN ORACLE MODEL AND A 'DO NOTHING' BASELINE.

Band-limited noise			System		
Central frequency (Hz)	Bandwidth (Hz)	Number of noisy bands	New system	Oracle model	Do nothing
656	175	1	68.0	65.3	58.7
1031	225	2	64.0	60.0	51.6
1265	325	3	46.2	45.3	28.9
2156	400	3	47.6	48.0	28.9

TABLE III

SPEAKER IDENTIFICATION ACCURACY (%) ON THE SPIDRE DATABASE, WITH VARIOUS TYPES OF REALISTIC NOISE CORRUPTION AT VARIABLE SNRS, AS A FUNCTION OF THE DURATIONS OF TRAINING AND TEST DATA, COMPARING THE NEW SYSTEM WITH THREE OTHER SYSTEMS.

System	Duration (s)		Noise type and SNR (dB)									
	Training	Testing	Clean	Restaurant			Street			Pop-song		
				10	15	20	10	15	20	10	15	20
New system	10	10	81.4	63.7	73.3	76.3	72.6	76.3	78.5	77.0	78.5	79.3
		5	80.7	60.0	71.9	74.1	67.4	76.3	79.3	70.4	75.6	76.3
	5	10	79.2	54.1	67.4	72.9	59.3	73.3	74.8	66.7	74.1	73.3
		5	71.9	46.7	63.7	70.4	56.3	68.1	76.3	60.0	71.9	70.4
Optimal feature only	10	10	82.2	45.2	62.2	74.1	43.0	62.2	75.6	68.9	76.3	80.7
		5	80.7	43.7	60.7	71.1	42.2	62.2	73.3	62.2	70.4	75.6
	5	10	82.2	39.3	55.6	68.1	41.5	55.6	69.6	59.3	71.1	77.0
		5	78.7	38.5	50.4	65.9	38.5	51.9	68.1	53.3	63.7	73.3
Multicondition only	10	10	81.5	51.9	66.7	77.0	65.9	74.8	82.2	69.6	79.3	80.7
		5	77.0	50.4	64.4	71.9	60.0	73.3	77.0	68.9	73.3	78.5
	5	10	72.6	41.5	51.9	63.7	55.6	68.1	74.1	60.0	67.4	71.1
		5	68.6	37.8	51.1	60.0	57.8	65.9	70.4	56.3	62.2	65.2
Do nothing	10	10	83.7	42.2	57.8	71.1	42.2	61.5	76.3	65.9	76.3	80.7
		5	81.5	40.7	54.1	68.1	41.5	58.5	74.1	62.2	73.3	75.6
	5	10	80.0	31.9	47.4	60.7	39.3	51.9	68.9	56.3	68.9	75.6
		5	78.5	34.1	43.0	58.5	37.8	51.9	64.4	51.1	65.9	71.1

TABLE IV

FACIAL IDENTIFICATION ACCURACY (%) ON THE AR DATABASE WITH A SINGLE TRAINING IMAGE PER PERSON, COMPARING THE NEW SYSTEM AGAINST OTHER SYSTEMS IN RECENT LITERATURE.

System	Clean	Sunglasses	Scarf
New system	100	89.6	94.8
J. Lin [24]	n/a	72.0	87.0
R. Akbari [23]	n/a	79.0	71.0
N.-S. Vu [40]	n/a	81.0	98.9
Z. Li [22]	n/a	77.3	89.9

TABLE V

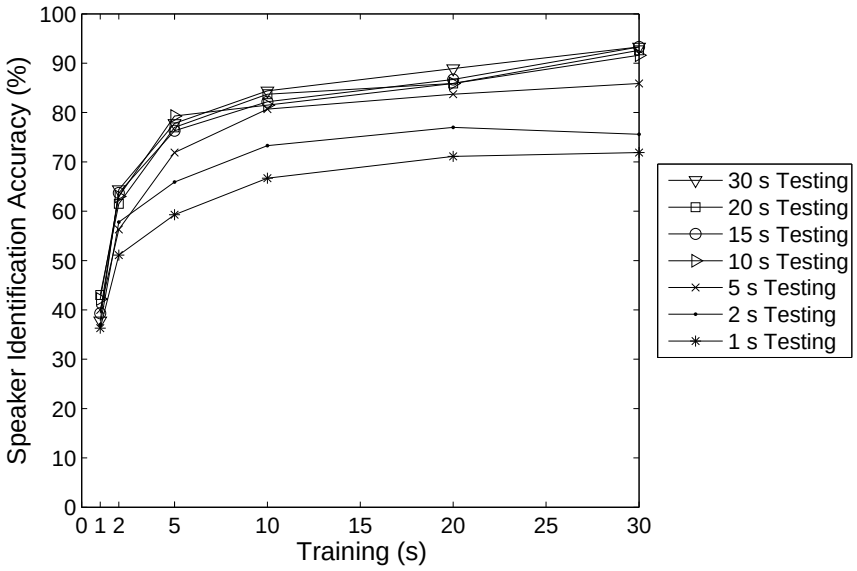
BIMODAL PERSON IDENTIFICATION ACCURACY (%) WITH LIMITED TRAINING SPEECH AND A SINGLE TRAINING FACIAL IMAGE, USING CLEAN TEST SPEECH AND OCCLUDED TEST IMAGES, AS A FUNCTION OF THE DURATIONS OF TRAINING AND TESTING SPEECH AND TYPE OF FACIAL OCCLUSION, COMPARING THE NEW SYSTEM AGAINST A 'DO NOTHING' MODEL.

System		New system						Do nothing					
	Testing (s)	10	5	10	5	10	5	10	5	10	5	10	5
Training (s)	Occlusion	Clean		Sunglasses		Scarf		Clean		Sunglasses		Scarf	
10		100	100	99.3	99.3	99.3	100	100	100	97.0	97.0	95.6	95.6
5		100	100	98.5	98.5	98.5	100	100	100	94.8	96.3	96.3	96.3

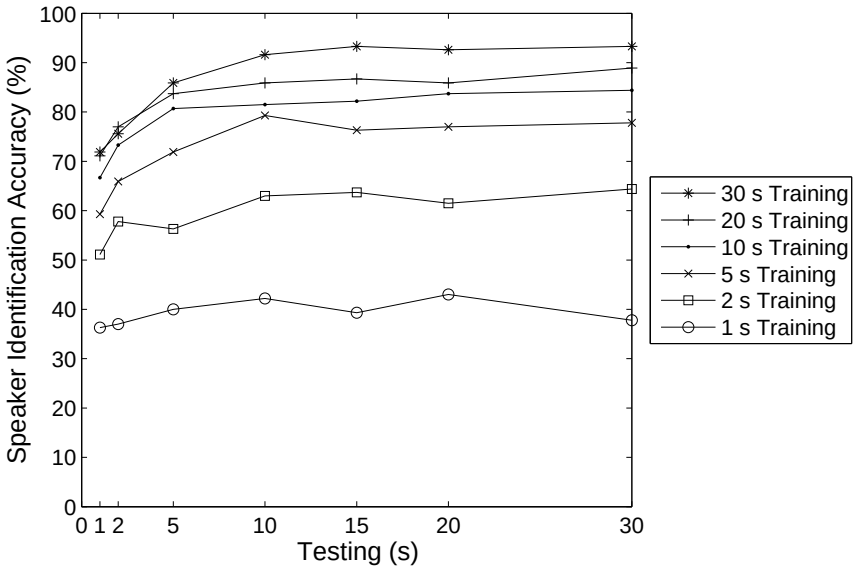
TABLE VI

BIMODAL PERSON IDENTIFICATION ACCURACY (%) WITH LIMITED TRAINING SPEECH AND A SINGLE TRAINING FACIAL IMAGE, USING NOISY TEST SPEECH AND OCCLUDED TEST IMAGES, AS A FUNCTION OF THE DURATIONS OF TRAINING AND TESTING SPEECH, TYPE OF ACOUSTIC NOISE AND SNR, AND TYPE OF FACIAL OCCLUSION, COMPARING THE NEW SYSTEM AGAINST A 'DO-NOTHING' MODEL.

System			New system				Do nothing			
	Testing (s)		10	5	10	5	10	5	10	5
Training (s)	Noise & SNR (dB)\Occlusion		Sunglasses		Scarf		Sunglasses		Scarf	
10	Restaurant	10	100	98.5	97.0	97.0	94.8	95.6	95.6	95.6
		15	100	100	97.8	98.5	96.3	96.3	95.6	95.6
		20	99.3	99.3	98.5	98.5	97.0	96.3	96.3	96.6
5		10	99.3	98.5	97.0	97.0	95.6	94.8	95.6	95.6
		15	98.5	100	97.8	97.0	96.3	97.0	95.6	96.3
		20	99.3	100	97.8	98.5	96.3	96.3	95.6	96.3
10	Street	10	100	100	97.0	97.8	96.3	97.8	95.6	95.6
		15	100	100	98.5	99.3	97.0	97.0	95.6	95.6
		20	99.3	99.3	98.5	99.3	96.3	97.0	95.6	95.6
5		10	100	100	97.0	97.0	96.3	97.0	94.8	95.6
		15	100	100	97.8	97.8	96.3	97.8	95.6	95.6
		20	100	100	98.5	100	95.6	97.0	95.6	95.6
10	Pop-song	10	99.3	100	97.8	98.5	97.0	96.3	95.6	95.6
		15	100	100	98.5	98.5	95.6	96.3	95.6	96.3
		20	100	99.3	99.3	99.3	96.3	96.3	95.6	97.0
5		10	100	100	97.8	97.8	97.0	97.0	95.6	95.6
		15	100	100	97.8	98.5	97.0	97.8	96.3	96.3
		20	98.5	99.3	98.5	99.3	96.3	96.3	95.6	96.3



(a)



(b)

Fig. 1. Speaker identification accuracy on the SPIDRE database. (a) Effect of training data duration. (b) Effect of test data duration.

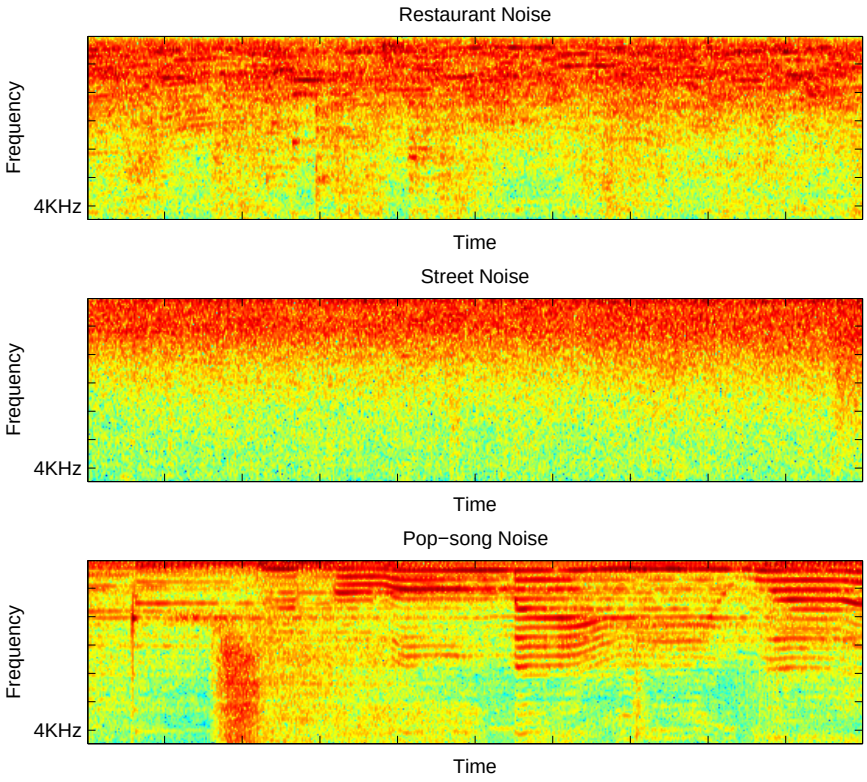


Fig. 2. Spectrograms showing 5 s samples from the three realistic noise types, restaurant, street and pop-song used to corrupt the test speech samples.

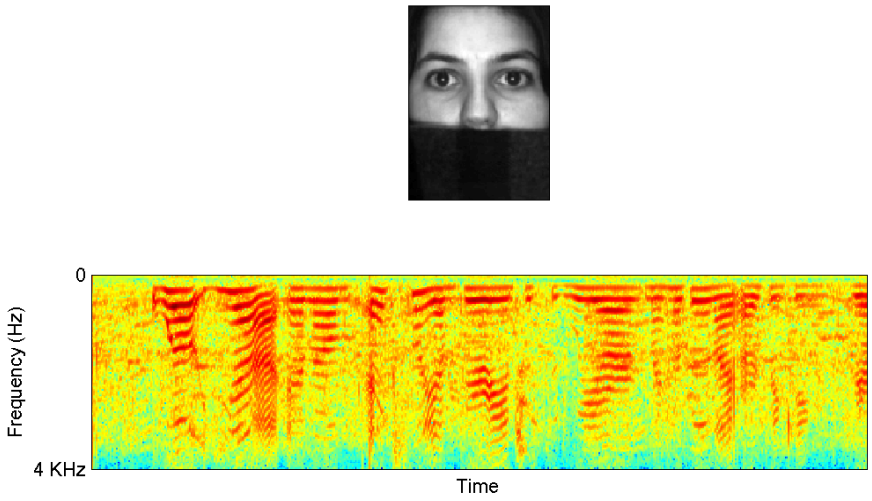


Fig. 3. An example of bimodal test data used in our identification system. Shown are an image of a subject from the AR face database with scarf occlusion, and a spectrogram with several seconds of speech data corrupted by restaurant noise with an SNR=10 dB.